融合扰动感知模型的孪生神经网络目标跟踪

李勇, 杨德东*, 韩亚君, 宋鹏
河北工业大学人工智能与数据科学学院, 天津 300130

摘要 针对全卷积孪生网络目标跟踪算法(Siamfc)在严重遮挡、旋转、光照变化、尺度变化等情况下容易出现跟踪失败的问题,提出了一种融合扰动感知模型的孪生神经网络目标跟踪算法。将孪生神经网络提取到的低层结构特征与高层语义特征进行有效融合,以提高特征的表征能力;利用模板自适应策略在线更新模板,以提高算法在遮挡和旋转等情况下跟踪的精确度。与此同时,将基于颜色直方图特征的扰动感知模型引入到算法中,通过加权融合的方式获得目标响应得分图,以此估计出目标的位置,并利用相邻帧尺度自适应策略估计出目标最佳尺度。为验证本文算法的效果,利用公开数据集测试所提算法性能,并与多种跟踪方法进行对比。实验结果表明:在 2015 目标跟踪标准测试数据集下本文所提算法总体跟踪精确度为 0.945,总体成功率为 0.929,相比 Siamfc 算法分别提高了 2.9%和 2.8%,在无人机航拍测试数据集中本文所提算法也具备较高的精确度与成功率,获得的跟踪效果良好。

关键词 机器视觉; 孪生神经网络; 扰动感知模型; 自适应模板; 特征融合
中图分类号 TP391.41 **文献标志码** A **doi**: 10.3788/AOS202040.0415002

Siamese Neural Network Object Tracking with Distractor-Aware Model

Li Yong, Yang Dedong*, Han Yajun, Song Peng
School of Artificial Intelligence, Hebei University of Technology, Tianjin 300130, *China*

Abstract Considering that the fully-convolutional siamese network algorithm for object tracking (Siamfc) algorithm is prone to tracking failure in cases such as heavy occlusion, rotation, illumination variation, scale variation, a siamese neural-network object-tracking algorithm with the distractor-aware model is proposed. First, the low-layer structural and high-layer semantic features were extracted from siamese networks; then, they were effectively fused to improve the representation ability of the feature. Second, the template adaptive strategy was used to update the template online to improve tracking accuracy in cases of occlusion and rotation. Simultaneously, the distractor-aware model based on color histogram features was introduced into the algorithm. The target response map was obtained by weighted fusion to estimate the position of the target while the adjacent frame scale adaptive strategy was used to estimate the optimal scale. To verify the effectiveness of the proposed algorithm, its performance was compared with those of various tracking methods on open-source datasets. Experimental results on the standard test dataset of the 2015th object tracking show that the overall tracking accuracy and success rate of the proposed algorithm are 0.945 and 0.929, which is 2.9% and 2.8% higher than those of the Siamfc algorithm, respectively. Further, the proposed algorithm performs with high accuracy and success rate in the aerial test dataset of an unmanned aerial vehicle (UAV).

Key words machine vision; siamese neural networks; distractor-aware model; adaptive template; feature fusion
OCIS codes 150.0150; 150.0155; 150.1135

1 引 言

视觉目标跟踪是计算机视觉中的一个基础研究课题,其应用非常广泛,包括人机交互、自动驾驶、交通监控、增强现实等方面[1-5]。常规跟踪任务是仅根据第一帧中目标的初始位置来估计图像序列中目标的移动轨迹。尽管在过去几十年视觉目标跟踪技术取得了巨大进步,但由于存在不可预测的外观变化,如光照变化、几何变形、部分遮挡、背景杂乱、快速运动等情况,现有视觉目标跟技术仍然面临着严峻的挑战。

近几年来,由于深度卷积网络具备强大的特征表征能力,逐渐被引入到视觉目标跟踪领域中,故大

收稿日期: 2019-08-26; **修回日期**: 2019-10-10; **录用日期**: 2019-11-06
基金项目: 河北省自然科学基金面上项目(F2017202009)
* **E-mail**: ydd12677@163.com

量基于深度卷积网络的视觉目标跟踪算法不断涌现。Wang 等[6]提出了全卷积网络视觉目标跟踪(FCNT)算法，发现了不同层次的卷积特征可以从不同的角度表征目标，并且通过观察发现视觉几何群网络(VGG-16)中有两层特征具备较好的互补性，通过融合这两层特征，有效抑制了跟踪过程中的跟踪器漂移现象，获得了较高的跟踪精度；Danelljan 等[7]提出了基于卷积特征的相关滤波视觉目标跟踪(DeepSRDCF)算法，使用预先训练的卷积网络提取目标特征，然后将这些特征输入到相关滤波器跟踪框架进行跟踪；Nam 等[8]提出了用于视觉跟踪的多域卷积神经网络(MDNet)算法，使用大量跟踪视频的手工标定真实值来对卷积神经网络进行预训练，学习到一个通用的目标检测器后，使用传统的检测跟踪框架进行目标跟踪。虽然这些跟踪算法取得了较好的效果，但由于要进行高维数据计算，所需的计算开销比较大，因此他们仍然无法实现实时跟踪。然而，目标跟踪研究者经过不懈努力，已经陆续提出了一些可以用于实时跟踪的深度神经网络跟踪器，Held 等[9]提出了深度回归网络视觉目标跟踪(GOTURN)算法，将前一帧获得的目标图像与当前帧搜索图像同时输入到卷积神经网络以输出级联特征，然后将级联特征输入全连接层以回归当前帧目标位置，由于网络结构相对简单，且没有设计模板在线更新功能，该算法实时性能良好；Tao 等[10]提出了孪生实例搜索视觉目标跟踪(SINT)算法，通过使用涵盖多种目标变化情况的大规模训练数据集离线训练模板匹配网络，在不应用任何模型更新，没有遮挡检测，没有跟踪器的组合，没有几何匹配的情况下实现了目标跟踪；Bertinetto 等[11]提出了基于全连接孪生网络的视觉目标跟踪(Siamfc)算法，使用孪生网络的完全卷积结构来计算搜索区域中每个位置的响应值，最高响应得分位置即目标位置。通过在跟踪基准测试集上测试 Siamfc 视觉目标跟踪算法性能，发现它可以达到实时跟踪要求，但是 Siamfc 算法也存在一定的缺陷，它不具备模板在线更新能力，仅使用高层输出的语义特征进行相关操作来获取目标位置，会导致算法在光照剧烈变化、严重几何形变、严重遮挡、背景杂乱、快速运动等复杂情况下跟踪失败。

针对上述情况，本文在 Siamfc 算法的基础上增加了模板在线更新模块，并将高层语义特征与低层结构特征进行加权融合以提高跟踪精度。与此同时，提取目标颜色特征，利用贝叶斯分类器来获得目标响应得分图，并将其与改进的全连接孪生网络跟踪算法响应得分图进行加权融合，以更好地处理目标形变和运动模糊问题。为验证本文算法效果，使用 2015 目标跟踪标准测试数据集[12]中视频序列和无人机航拍测试数据集[13]中视频序列测试本文所提算法性能。

2　Siamfc 跟踪算法

Siamfc 跟踪算法的关键在于它离线训练一个分支权重共享的孪生网络，然后使用此训练好的网络进行在线跟踪；此外，它通过使用模板特征与输入图像特征进行相关性匹配方法，获得相关性响应得分图，通过定位最大响应位置找到目标位置。由于其实现原理和网络结构相对简单，其算法复杂度较低，因此 Siamfc 算法实时性良好。

2.1　网络结构

孪生神经网络由两个相同的分支组成，这两个分支相互共享参数，通过将模板特征和当前帧图像特征输入互相关层进行相关操作，获得候选目标与模板的相似度得分，以实现目标跟踪，其网络结构示意图如图 1 所示。

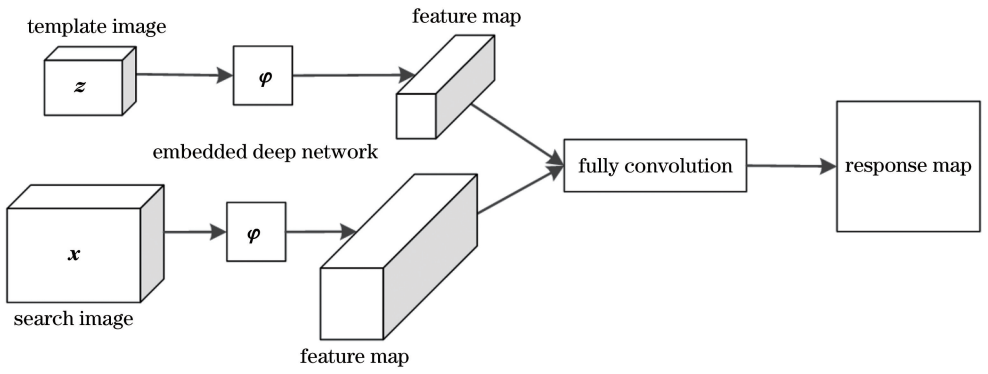

图 1　孪生神经网络示意图

Fig. 1　Schematic diagram of siamese neural network

对于 Siamfc 算法而言，其关键点就是学习一个相似度函数，然后利用它计算两个分支提取到的模板特征和输入图像特征的相似度，以获得各像素点相似度得分。如果该像素属于目标则返回高分，如果不属于目标则返回低分。而全连接孪生神经网络所使用的相似度函数为

$$\boldsymbol{f}(\boldsymbol{z},\boldsymbol{x})=\boldsymbol{\varphi}(\boldsymbol{z})*\boldsymbol{\varphi}(\boldsymbol{x})+\boldsymbol{b}, \tag{1}$$

式中：$\boldsymbol{x}$ 是输入的搜索图像；$\boldsymbol{z}$ 是模板图像；变换 $\boldsymbol{\varphi}$ 为卷积嵌入函数，它对应孪生网络的特征提取阶段；$*$ 表示互相关运算；$\boldsymbol{b}$ 表示在每个位置取值为 $\boldsymbol{b}\in\mathbb{R}^{n\times n}$ 的偏差信号，$\mathbb{R}^{n\times n}$ 表示 $n\times n$ 的实数矩阵，$\mathbb{R}$ 表示实数集，n 表示矩阵维度。$\boldsymbol{f}(\boldsymbol{z},\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 与 $\boldsymbol{z}$ 的相似度得分，其输出是一张响应得分图，该响应图中得分最高的位置即目标位置。

2.2　损失函数

模型训练时，必不可少的是损失函数，这是因为利用最小化损失函数来获取最优化模型参数是一个最有效的途径。Siamfc 算法为了构造有效的损失函数，对用于训练的搜索图像区域划分为正负样本，即将距离目标中心一定范围内的点作为正样本，超过这个范围的点规定为负样本。按照这样构造的损失函数实质上是一个典型的二分类问题，故使用平均二分类逻辑损失函数可以有效表征其损失，即

$$L(y,v)=\frac{1}{N_{\boldsymbol{G}}}\sum_{\mu\in\boldsymbol{G}}l(y[\mu],v[\mu]), \tag{2}$$

式中：$y[\mu]\in[+1,-1]$ 表征真实的样本类别；$v[\mu]$ 是每个搜索位置 μ 的响应得分；$\boldsymbol{G}$ 表示最终生成的响应得分矩阵，$N_{\boldsymbol{G}}$ 表示该得分矩阵中数值个数；l 为单个点逻辑损失，表达式为

$$l(y[\mu],v[\mu])=\log[1+\exp(-y[\mu]v[\mu])]。 \tag{3}$$

当该样本点为正样本时，$y[\mu]=+1$，若 $v[\mu]$ 取较大值，此时 $l(y[\mu],v[\mu])$ 将获得一个较小值，说明跟踪正确时损失较小；当该样本为负样本时，$y[\mu]=-1$，若 $v[\mu]$ 仍取较大值，此时 $l(y[\mu],v[\mu])$ 将获得一个较大值，说明跟踪错误时损失函数将取得较大值。这样通过最小化损失函数就能获得网络模型最优参数。在 Siamfc 算法中采用了随机梯度下降法(SGD)来最小化损失函数以获得最优化模型参数 θ，这里的模型参数 θ 指代的是孪生神经网络的各层参数。

3　本文算法

本文在 Siamfc 算法的基础上提出融合扰动感知模型的孪生神经网络目标跟踪算法，图 2 为本文算法的跟踪框架图。

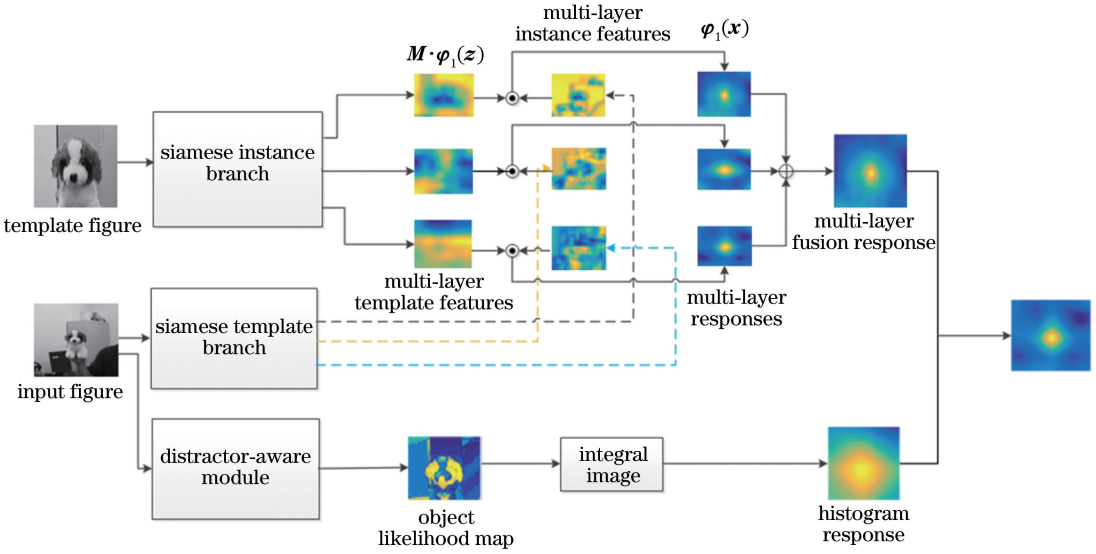

图 2　本文算法框架图

Fig. 2　Algorithmic framework diagram in this paper

3.1　基于孪生神经网络的目标跟踪

3.1.1　多层卷积特征融合

孪生神经网络实质上是由两个共享权重的卷积神经网络所构成的，其通过对称的卷积神经网络分别提取模板和待搜索图像深层次卷积特征，以进行模板匹配来获得最终响应得分图，由于卷积特征具备较强的表征能力，因此其用于视觉目标跟踪时具有得天独厚的优势。但是不同层卷积特征所表征的目标信息有差异，如图 3 各层卷积特征可视化结果所示，高层次卷积特征更多地表征目标的语义信息，

低层次特征更多地表征目标的结构信息，而视觉目标跟踪任务不仅需要通过判别语义信息来区分不同对象，还需要结合结构信息来精确定位目标位置，因此对多层卷积特征进行融合有助于提高算法的跟踪精确度。

本文算法对多层卷积特征响应进行融合以得到置信度较高的多层特征融合响应图，具体表示形式为

$$\boldsymbol{f}_t=\sum_{s=i}w_t^s\boldsymbol{f}_t^s, \tag{4}$$

式中：w_t^s 表示第 t 帧第 s 层特征响应得分图的权重；$\boldsymbol{f}_t$ 表示第 t 帧的多层目标特征响应融合后的响应得分图；$\boldsymbol{f}_t^s$ 表示第 t 帧第 s 层特征响应得分图；这里 i 为需要融合的特征层序号，取值为 $i=2,4,5$。

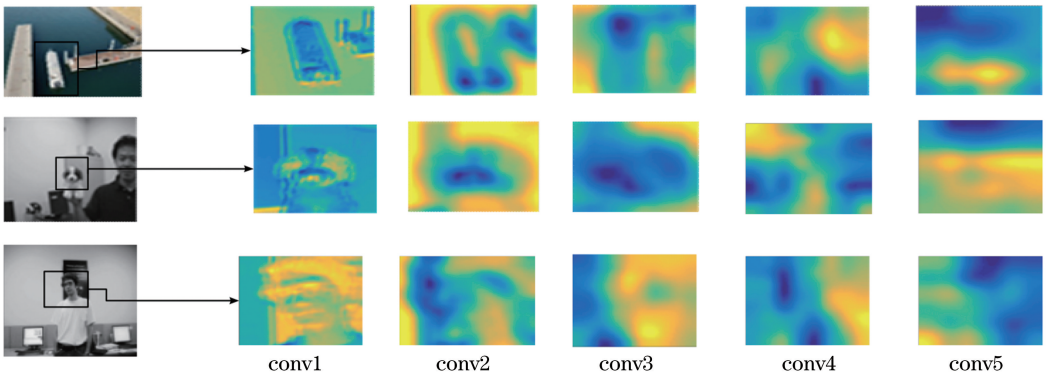

图 3　各层卷积特征图可视化结果

Fig. 3　Visualization of each layer's convolutional feature map

3.1.2　模板自适应策略

由于 Siamfc 算法以第一帧输入图像标定真值区域作为匹配模板，在线跟踪时没有对模板进行在线更新，因此该算法在严重遮挡、剧烈形变等情况下容易跟丢目标。本文算法引入文献[14]中的在线模板更新方法进行模板更新以计算相关性响应得分图，表达式为

$$\boldsymbol{f}_t(\boldsymbol{z},\boldsymbol{x})=\boldsymbol{M}_{t-1}\cdot\boldsymbol{\varphi}_1(\boldsymbol{z})*\boldsymbol{\varphi}_t(\boldsymbol{x})+\boldsymbol{b}, \tag{5}$$

式中：$\boldsymbol{M}_{t-1}=\mathscr{F}^{-1}\left\{\dfrac{\boldsymbol{\varphi}_1^*(\boldsymbol{z})\odot\boldsymbol{\varphi}_{t-1}(\boldsymbol{z})}{\boldsymbol{\varphi}_1^*(\boldsymbol{z})\odot\boldsymbol{\varphi}_1(\boldsymbol{z})+\lambda_\beta}\right\}$ 表征模板变化程度，$\boldsymbol{\varphi}_1^*(\boldsymbol{z})$ 为第一帧模板特征共轭转置，$\odot$ 表示矩阵元素点乘操作，λ_β 是正则化参数，其目的是防止过拟合，$\mathscr{F}^{-1}\{\cdot\}$ 表示傅里叶逆变换。

3.2　扰动感知建模

颜色直方图特征尽管对颜色比较敏感，但是在形变和相似目标干扰等场景下具有很好的跟踪效果，因此将其引入到算法中能在一定程度上增强跟踪鲁棒性。而且作为算法分支，它有效建立了能预先识别潜在干扰区域的扰动感知模型，能够大大减少漂移现象发生的频次。

3.2.1　建立目标-扰动模型

为了有效地区分目标前景与背景，所提算法利用基于灰度直方图的贝叶斯分类器对目标的灰度信息进行建模，对于给定的目标前景 $\boldsymbol{O}$ 和背景 $\boldsymbol{B}$，根据贝叶斯公式可得到搜索区域内的像素 x 属于目标前景的概率分布为

$$\boldsymbol{P}(x\in\boldsymbol{O}\mid\boldsymbol{O},\boldsymbol{B},h_x)\approx\frac{\boldsymbol{P}(h_x\mid x\in\boldsymbol{O})\boldsymbol{P}(x\in\boldsymbol{O})}{\sum\limits_{\Omega\in\{\boldsymbol{O},\boldsymbol{B}\}}\boldsymbol{P}(h_x\mid x\in\Omega)\boldsymbol{P}(x\in\Omega)}, \tag{6}$$

式中：h_x 表示像素点 x 的亮度值在灰度区间 h 的范围内。概率分布可以使用灰度直方图来表征，即：$\boldsymbol{P}(h_x\mid x\in\boldsymbol{O})\approx\boldsymbol{H}_{\boldsymbol{O}}^I(h_x)/|\boldsymbol{O}|$，$\boldsymbol{P}(h_x\mid x\in\boldsymbol{B})\approx\boldsymbol{H}_{\boldsymbol{B}}^I(h_x)/|\boldsymbol{B}|$，其中 $\boldsymbol{H}_\Omega^I(h_x)$ 表示在图像 $\boldsymbol{I}$ 中区域 Ω 的第 h 个灰度区间直方图。那么(6)式化简为

$$\boldsymbol{P}(x\in\boldsymbol{O}\mid\boldsymbol{O},\boldsymbol{B},h_x)=\begin{cases}\dfrac{\boldsymbol{H}_{\boldsymbol{O}}^I(h_x)}{\boldsymbol{H}_{\boldsymbol{O}}^I(h_x)+\boldsymbol{H}_{\boldsymbol{B}}^I(h_x)}, & \text{if}\quad\boldsymbol{V}(x)\in\boldsymbol{V}(\boldsymbol{O}\cup\boldsymbol{B})\\ 0.5, & \text{otherwise}\end{cases}, \tag{7}$$

式中：$\boldsymbol{V}(\cdot)$ 为像素亮度值计算函数。(7)式所建立的模型即为目标的前景-背景模型，通过这样建模即可将目标与背景有效区分。但是由于背景中还可能存在与目标颜色相似的区域，故建立目标-扰动模型为

$$\boldsymbol{P}(x\in\boldsymbol{O}\mid\boldsymbol{O},\boldsymbol{D},h_x)=\begin{cases}\dfrac{\boldsymbol{H}_{\boldsymbol{O}}^I(h_x)}{\boldsymbol{H}_{\boldsymbol{O}}^I(h_x)+\boldsymbol{H}_{\boldsymbol{D}}^I(h_x)}, & \text{if}\quad\boldsymbol{V}(x)\in\boldsymbol{V}(\boldsymbol{O}\cup\boldsymbol{D})\\ 0.5, & \text{otherwise}\end{cases}, \tag{8}$$

式中：$\boldsymbol{D}$ 为与目标颜色相似的环境区域。将上述两个模型整合得到最终的判别式目标模型为

$$\boldsymbol{P}(x\in\boldsymbol{O}\mid h_x)=\lambda_p\boldsymbol{P}(x\in\boldsymbol{O}\mid\boldsymbol{O},\boldsymbol{D},h_x)+(1-\lambda_p)\boldsymbol{P}(x\in\boldsymbol{O}\mid\boldsymbol{O},\boldsymbol{B},h_x),\tag{9}$$

式中：λ_p 为预设的权重参数。为适应形变和光照的变化，使用下述更新模型，即

$$\boldsymbol{P}_{1:t}(x\in\boldsymbol{O}\mid h_x)=\eta_p\boldsymbol{P}(x\in\boldsymbol{O}\mid h_x)+(1-\eta_p)\boldsymbol{P}_{1:t-1}(x\in\boldsymbol{O}\mid h_x),\tag{10}$$

式中：η_p 为学习率，下标 $1:t$ 表示从第 1 帧到第 t 帧。

3.2.2　颜色直方图得分

在目标定位阶段，以上一帧目标中心位置为中心，选取搜索框 $\boldsymbol{W}\in\boldsymbol{I}$，$\boldsymbol{I}$ 表示当前帧输入图像，根据(9)式可计算出该搜索框内每个像素 x 属于目标 $\boldsymbol{O}$ 的概率分布为

$$\boldsymbol{P}_t(x\in\boldsymbol{O})=\boldsymbol{P}_{1:t}(x\in\boldsymbol{O}\mid h_x),\tag{11}$$

本文算法直方图得分可通过对上述 $\boldsymbol{P}_t(x\in\boldsymbol{O})$ 的积分图取平均得到，表达式为

$$\boldsymbol{f}_{\mathrm{hist}}(\boldsymbol{I})=\frac{1}{|\boldsymbol{W}|}\sum_{x\in\boldsymbol{W}}\boldsymbol{P}_t(x)。\tag{12}$$

式中：hist 表示为直方图。

3.3　目标定位

分别算出基于孪生神经网络跟踪响应得分图 $\boldsymbol{f}_t(\boldsymbol{z},\boldsymbol{x})$ 和基于彩色特征跟踪的响应得分图 $\boldsymbol{f}_{\mathrm{hist}}(\boldsymbol{I})$，然后使用线性求和的方式进行融合，即可得到最终的响应得分图为

$$\boldsymbol{f}(\boldsymbol{I})=\gamma\boldsymbol{f}_t(\boldsymbol{z},\boldsymbol{x})+(1-\gamma)\boldsymbol{f}_{\mathrm{hist}}(\boldsymbol{I}),\tag{13}$$

式中：γ 为融合扰动感知模型参数。通过寻找最终响应得分图得分最大的位置来进行目标定位，即

$$\boldsymbol{Y}_t=\mathrm{argmax}\,\boldsymbol{f}(\boldsymbol{I}),\tag{14}$$

式中，最终响应得分图 $\boldsymbol{f}(\boldsymbol{I})$ 中得分最大的位置坐标 $\boldsymbol{Y}_t$ 与需要寻找的目标中心位置坐标相对应。通过这种方式即可实现目标的定位。

3.4　尺度自适应

为了使得算法在能够应对目标尺度变化场景的同时不会导致算法时间复杂度剧烈提升，本文算法对每帧搜索图像进行三次尺度估计，首先提取候选区域的三个尺度，即

$$\boldsymbol{x}_t^i=s_t^i\boldsymbol{x}_1^i,\tag{15}$$

式中：下标 t 表示搜索图像的帧序号；上标 i 表示该帧图像的候选尺度序号；s 表示尺度因子；第 t 帧第 i 个候选尺度因子 $s_t^i=s_{t-1}^{\mathrm{opt}}\Gamma$，其中 s_{t-1}^{opt} 为第 $t-1$ 帧的最优尺度因子，Γ 是尺度变化率。最终通过比较各尺度响应得分图的最大值来获得最佳响应效果和最优尺度因子。

4　实验结果分析

4.1　实验环境及参数设定

本文算法运行平台配置为：CPU 为 6 核 3.7 GHz Intel I7，内存为 16G，显卡为 GTX1080TI，操作系统为 64 位 windows10，所采用编程环境为 Matlab2015b。算法所涉及的参数设定如下：扰动感知模型预设的权重参数 $\lambda_p=0.5$，扰动感知模型学习率 $\eta_p=0.1$，融合扰动感知模型参数 $\gamma=0.3$；所用融合层 i 分别是第二层、四层和第五层，其所对应融合权重 w_t^s 分别为 0.6、0.3、0.1，以上参数均参照文献[15]中的方法，通过对参数进行大量调试来获得。尺度变化率 $\Gamma=[0.9639,1,1.0375]$，该参数参照文献[16]提供的参数进行设定。

4.2　利用 2015 目标跟踪标准测试集测试算法性能

为验证本文所提算法的效果，利用了 2015 目标跟踪标准测试数据集中的 10 组视频序列进行测试，所用视频序列涵盖了遮挡、旋转、形变、光照变化、尺度变化等属性，各组视频序列的长度、分辨率及属性详见表 1。同时将本文算法所测试结果与 Siamfc[10]、DSiamM[14]、ASLA[17]、TLD[18]、MEEM[19]、MUSTER[20]、IVT[21] 等 7 种当前比较流行的算法进行对比。

对比过程中所用评价指标为目标中心位置误差、精确度、重叠率、成功率等。其中：目标中心位置误差指的是实际目标跟踪框中心位置与人工标定准确目标中心位置之间的平均欧氏距离。精确度指的是目标中心位置误差值小于某一阈值的帧数占总帧数的比值，这里阈值设置为 20 pixel。重叠率指的是实际跟踪框与标定跟踪框之间重叠面积占两跟踪框并集面积的比例。成功率指的是重叠率大于一定阈值的帧数与总帧数的比值，这里取该阈值为 0.5。图 4 所示为本文算法与其他算法在所测试序列中的总体精确度与成功率图，从图中可以看出本文算法总体精确度为 0.945，成功率为 0.929，相比基础算法 Siamfc 分别提升了 2.9% 和 2.8%。

为分析本文算法鲁棒性，详细测试了 8 种算法在这 10 组视频序列中的跟踪误差，如表 2 所示，从表中可以看出，本文算法在 8 组序列中跟踪误差性能均排名前 3，且在大多数序列中其跟踪误差较 Siamfc 算法的跟踪误差低，这说明本文所提算法不仅在总体成功率和精确度上获得了良好性能，而且在复杂场景中也具备较好的鲁棒性。

表 1　10 组视频属性

Table 1　Ten sets of video attributes

Video sequence	Length / frame	Resolution ratio / (pixel×pixel)	Characteristic
David2	537	320×240	In-plane rotation, out-of-plane rotation
Faceocc1	892	352×288	Occlusion
Faceocc2	812	320×240	Illumination variation, occlusion, in-plane rotation, out-of-plane rotation
Subway	175	352×288	Occlusion, deformation, background clutter
Freeman1	326	360×240	Scale variation, in-plane rotation, out-of-plane rotation
MountainBike	228	640×360	Out-of-plane rotation, in-plane rotation, background clutter
Dog1	1350	320×240	Scale variation, in-plane rotation, out-of-plane rotation
CarScale	252	640×272	Scale variation, occlusion, fast motion, in-plane rotation, out-of-plane rotation
Football	362	624×352	Occlusion, in-plane rotation, out-of-plane rotation, background clutter
Basketball	725	576×432	Illumination variation, out-of-plane rotation, occlusion, deformation, background clutter

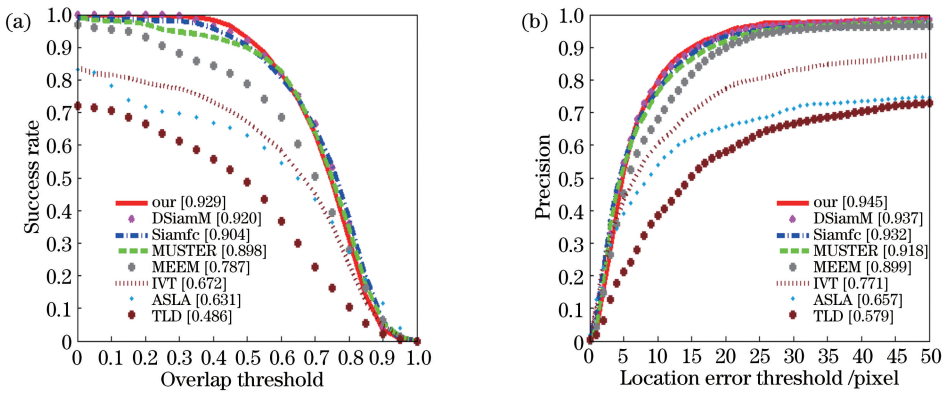

图 4　8 种跟踪算法的成功率图和精确度图。(a)成功率;(b)精确度

Fig. 4　Precision plots and success plots of the eight trackers. (a) Success rate; (b) accuracy

表 2　跟踪算法在 10 组视频序列中跟踪误差

Table 2　Tracking errors of tracking algorithms in ten video sequences

Sequence	Our	Siamfc	DSiamM	ASLA	TLD	MEEM	MUSTER	IVT
MountainBike	5.6199	6.1406	5.7915	8.9727	213.3278	13.0037	8.12	7.416
Faceocc1	10.1931	11.9656	11.4831	77.8108	27.3678	16.9904	14.2932	17.8346
Freeman1	5.9435	6.6078	6.039	104.8774	39.6988	11.3029	8.6361	11.7283
Subway	2.4955	3.254	2.9104	137.6901	159.0114	4.1169	2.2211	130.2318
Football	5.2967	6.7392	5.0698	15.3724	14.2587	5.1423	14.7789	14.8367
CarScale	15.7498	15.318	18.4334	24.9002	50.3495	67.2993	18.6758	11.7225
Basketball	10.4241	22.7174	10.658	82.6266	268.7569	4.2104	4.8487	106.9015
Faceocc2	10.1346	10.7052	10.1493	19.5059	12.2779	10.5872	5.8895	7.1397
Dog1	5.0088	3.004	3.5086	5.8068	4.1903	6.1053	4.0696	4.0764
David2	3.7716	2.8061	3.007	1.5874	4.9788	1.863	1.9849	1.6066

为了更直观地分析所提算法，分别记录了各个算法在遮挡、旋转、光照变化、尺度变化等情况下的实际跟踪效果（如图 5 所示）。

图 5　8 种算法实际效果图。(a) Faceocc1；(b) subway；(c) football；(d) freeman1；(e) dog；(f) carScale；(g) mountainBike；(h) david2；(i) faceocc2；(j) basketball

Fig. 5　Actual results of eight algorithms. (a) Faceocc1; (b) subway; (c) football; (d) freeman1; (e) dog1; (f) carScale; (g) mountainBike; (h) david2; (i) faceocc2; (j) basketball

遮挡情况下各算法性能对比分析：在 faceocc1 视频序列的第 487 帧图像中人脸被书本遮挡时，ASLA 算法由于使用单一的像素特征进行跟踪，其模型表征能力较弱，跟踪目标发生漂移，导致跟踪失败；而 TLD 算法由于其分类器分类能力较弱，也出现了跟踪失败情况。在 subway 序列的第 52 帧、第 72 帧图像中目标在背景杂乱环境下被遮挡时，不少算法由于特征表征能力弱或缺乏模板更新模块，跟踪失败，例如 IVT、ASLA 等算法。而本文所提算法由于采用多特征融合的扰动感知跟踪策略，其模型表征能力强，且使用模板更新方法使算法抗遮挡能力进一步增强，因此本文算法仍能较好地跟踪目标，特别是在 football 序列的第 322 帧图像中可以明显看出，由于运动员在背景杂乱环境下不仅被遮

挡，还出现了相似目标干扰，其他算法中的跟踪目标均出现漂移现象，仅本文算法能准确跟踪目标。

旋转情况下各算法性能对比分析：在 david2 视频序列中跟踪目标所要面临的挑战主要是平面内旋转和平面外旋转，不少算法由于所用特征表征能力较弱未能精准跟踪目标，而本文所提算法以及多数深度学习式跟踪算法均能较好地跟踪目标。在 MountainBike 的第 91 帧、第 197 帧图像中，ASLA、IVT 等传统跟踪算法由于目标所处场景相对复杂，跟踪目标出现了一定程度的漂移，而本文所提算法由于特征表征能力较强，以及使用了尺度自适应策略，故仍能准确跟踪目标。

光照变化情况下各算法性能对比分析：在 faceocc2 序列的第 275 和第 337 帧图像中可以明显看到，目标所处环境的光照亮度发生了改变，这时算法仅利用单一的颜色特征或直方图特征跟踪目标，很可能产生目标跟踪漂移现象，例如该序列中 MEEM 算法跟踪的目标就出现了轻度漂移。在 basketball 序列中的第 660 帧、697 帧和第 725 帧图像中，篮球运动员在移动过程中的位置发生改变，光照亮度发生变化，虽然基础算法 Siamfc 使用了深度特征进行目标跟踪，但是由于缺乏深浅层特征之间的融合，跟踪失败，而本文算法仍能精确跟踪目标。

尺度变化情况下各算法性能对比分析：在 freeman1 视频序列的第 185 帧图像和 dog1 视频序列的第 1028 帧图像中，由于目标离摄像镜头的距离改变导致目标尺度发生了变化，因此 MEEM、MUSTER 等跟踪算法由于缺乏相应的尺度自适应机制，在此目标尺度变化场景下跟踪漂移现象出现，最终导致跟踪失败。在 carScale 视频序列的第 171 帧、第 237 帧和第 252 帧图像中汽车发生了显著尺度变化，这时候本文所提跟踪算法仍能较好地跟踪目标，对汽车尺度变化具有一定的自适应能力。

4.3　利用航拍数据集测试算法性能

为进一步验证本文算法的有效性，从无人机航拍数据集 UAV123 中选取 10 组颇具挑战性的航拍视频序列测试本文算法效果，表 3 所示为该 10 组航拍视频序列的属性，其新引入了纵横比变化、视角变化、相机移动、相似目标等新属性。图 6 所示为本文算法与 Siamfc[10]、DSiamM[14]、ASLA[17]、TLD[18]、MEEM[19]、MUSTER[20]、IVT[21] 等 7 种算法在 10 组航拍视频序列中的跟踪成功率与精确度图，可以看出本文所提算法精确度为 0.963，成功率为 0.673，其较 Siamfc 算法分别提升了 25.1% 和 26%，可知在此测试视频序列上，本文所提算法跟踪效果较好，相对于基础算法提升幅度十分明显，且在 8 种算法中总体精确度和成功率都排名第一。

表 3　10 组航拍视频序列的属性

Table 3　Ten sets of aerial video sequence attributes

Video sequence	Length / frame	Resolution ratio / (pixel×pixel)	Characteristic
Wakeboard4	233	1280×720	Scale variation, aspect ratio change, viewpoint change
Wakeboard10	157	1280×720	Scale variation, low resolution
Boat1	301	1280×720	Scale variation
Boat2	267	1280×720	Scale variation
Boat6	269	1280×720	Scale variation
Boat9	467	1280×720	Scale variation, aspect ratio change, low resolution, partial occlusion, viewpoint change
Building1	157	1280×720	
Truck3	179	1280×720	Low resolution, partial occlusion, background clutter
Car4	449	1280×720	Occlusion, aspect ratio change, low resolution, partial occlusion, camera motion, similar object
Car5	249	1280×720	Scale variation

表 4 记录了各算法在单个视频序列中的精确度得分，表 5 所示为各算法在单个视频序列中成功率得分，从两个表中数据可以看出本文算法在各个测试视频序列中精确度和成功率较大多数对比算法高，算法的鲁棒性良好。图 7 记录了在航拍视频序列下各算法的实际跟踪效果，从图中可以看出本文

算法在应对尺度变化、纵横比变化、视角变化、相机移动、相似目标、部分遮挡等诸多挑战上表现出了相当不错的效果。通过在无人机航拍视频序列下对算法总体精确度和成功率、单个视频序列下精确度和成功率、算法实际跟踪效果等诸多方面进行对比，可以发现本文所提算法精确度和成功率较高，算法鲁棒性良好。

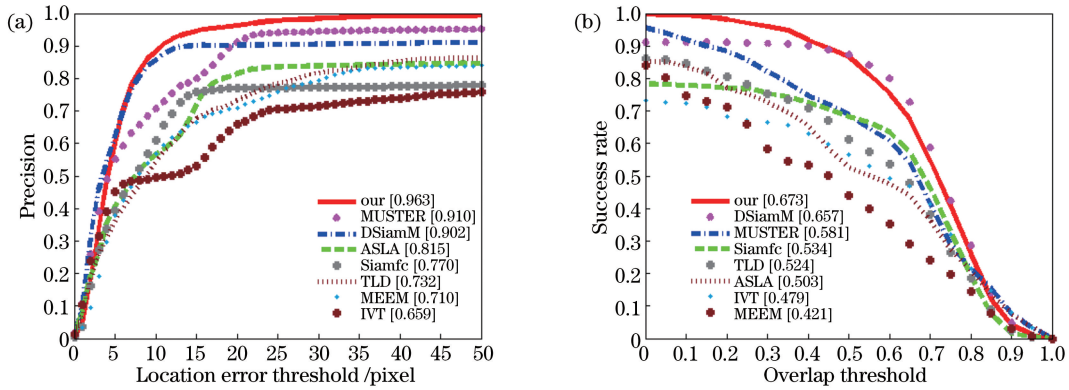

图 6　各种跟踪算法在航拍视频序列中的成功率图和精确度图。(a)精确度；(b)成功率

Fig. 6　Success rate and accuracy of various tracking algorithms in aerial video sequences. (a) Accuracy; (b) success rate

表 4　跟踪算法在 10 个视频序列中的精确度

Table 4　Accuracy of tracking algorithms in ten video sequences

Sequence	Our	MUSTER	DSiamM	ASLA	Siamfc	TLD	MEEM	IVT
Wakeboard4	0.751	0.549	0.588	0.004	0.305	0.077	0.597	0.004
Wakeboard10	1.000	1.000	1.000	0.917	1.000	1.000	1.000	0.248
Boat1	1.000	0.841	1.000	0.990	1.000	0.498	0.658	0.957
Boat2	1.000	1.000	1.000	1.000	1.000	0.397	1.000	1.000
Boat6	0.933	0.892	0.914	0.955	0.922	0.885	0.818	0.981
Boat9	0.953	0.914	0.522	0.829	0.972	0.473	0.469	0.203
Building1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Truck3	1.000	1.000	1.000	1.000	0.207	1.000	1.000	1.000
Car4	0.989	0.998	0.998	0.457	0.296	0.998	0.296	0.450
Car5	1.000	0.908	1.000	1.000	1.000	0.996	0.265	0.743

表 5　跟踪算法在 10 个视频序列中的成功率

Table 5　Success rate of tracking algorithms in ten video sequences

Sequence	Our	MUSTER	DSiamM	ASLA	Siamfc	TLD	MEEM	IVT
Wakeboard4	0.434	0.312	0.363	0.009	0.185	0.029	0.348	0.010
Wakeboard10	0.567	0.396	0.631	0.365	0.552	0.437	0.333	0.146
Boat1	0.730	0.731	0.722	0.529	0.740	0.595	0.376	0.612
Boat2	0.748	0.745	0.754	0.773	0.745	0.624	0.618	0.817
Boat6	0.786	0.339	0.774	0.602	0.763	0.346	0.329	0.618
Boat9	0.526	0.332	0.325	0.273	0.534	0.201	0.071	0.116
Building1	0.816	0.803	0.830	0.764	0.742	0.737	0.781	0.793
Truck3	0.613	0.794	0.702	0.832	0.132	0.753	0.694	0.787
Car4	0.756	0.635	0.706	0.362	0.227	0.785	0.244	0.369
Car5	0.757	0.721	0.766	0.521	0.722	0.729	0.412	0.526

图 7　在航拍视频序列下的算法实际效果对比

Fig. 7　Actual effect of algorithms in aerial video sequence

4.4　算法实时性评估

为验证算法的实时性能，在利用 2015 目标跟踪标准测试数据集测试算法精确度与成功率等性能的同时记录了各算法的平均速度，如表 6 所示，从表中数据可以发现 Siamfc 算法速度比所提算法的速度快，这是由于本文算法引入了模型自适应更新和多层特征融合等策略所致。而所提算法中孪生网络模块与扰动感知模块在融合前为并行计算，所以算法总体速度主要取决于运行速度较慢的那个模块。经过测定发现扰动感知模块单支运行平均速度为 49.53 frame/s，所以所提算法的速度主要取决于孪生网络模块，但由于所提算法仅进行三次尺度估计且其较 DSiamM 所使用的背景抑制方法而言，所提算法直接采用扰动感知模块中的扰动判别方案增强了抗背景干扰的能力，这就使得在保证算法精度获得提升的同时，时间复杂度较 DSiamM 算法低，其速度较 DSiamM 算法快。在实际场景应用中，所提算法可以较好满足实时跟踪要求。

表 6　算法平均速度对比

Table 6　Average speed comparison of the algorithms

Algorithm	Our	Siamfc	MUSTER	DSiamM	ASLA	TLD	MEEM	IVT
Speed /(frame · s^{-1})	37.2	56.1	3.7	23.9	8.1	27.8	9.9	39.9

5　结　　论

本文提出融合扰动感知模型的孪生神经网络目标跟踪算法，针对不同层次卷积特征所包含目标信息的差异性，将多层卷积特征进行有效融合，提高了算法的精确度；针对 Siamfc 算法仅使用第一帧作为模板搜索目标，其缺乏模板自适应更新易导致算法在严重遮挡和旋转等情况下跟踪失败的问题，所提算法使用孪生网络模板自适应策略在线更新模板，提高了算法在严重遮挡和旋转等情况下跟踪目标的成功率。同时，本文算法融合了扰动感知模型，提高了算法在复杂场景下进行目标跟踪的鲁棒性。最后使用相邻帧尺度自适应策略进行尺度估计，提高了算法在尺度变化情况下的跟踪性能。基于利用 2015 目标跟踪标准测试数据集中视频序列测试本文算法效果，发现所提算法总体跟踪精确度为

0.945，总体成功率为 0.929，相比 Siamfc 算法分别提高了2.9%和 2.8%；同时在无人机航拍数据集测试中，所提算法也表现良好。通过实验验证可得，本文所提算法在严重遮挡、旋转、光照变化、尺度变化等情况下能够较好地跟踪目标，具有一定的研究价值。

参 考 文 献

[1] Liu W J, Sun H, Jiang W T. Correlation filter tracking algorithm for adaptive feature selection[J]. Acta Optica Sinica, 2019, 39(6): 0615004.
刘万军, 孙虎, 姜文涛. 自适应特征选择的相关滤波跟踪算法[J]. 光学学报, 2019, 39(6): 0615004.

[2] Mao N, Yang D D, Li Y, et al. Spatial regularization correlation filtering tracking via deformable diversity similarity[J]. Acta Optica Sinica, 2019, 39(4): 0415002.
毛宁, 杨德东, 李勇, 等. 基于形变多样相似性的空间正则化相关滤波跟踪[J]. 光学学报, 2019, 39(4): 0415002.

[3] Zhang B, Jiang F B, Liu G. Context-aware tracking based on a visual saliency and perturbation model[J]. Optics and Precision Engineering, 2018, 26(8): 2112-2121.
张博, 江沸菠, 刘刚. 利用视觉显著性和扰动模型的上下文感知跟踪[J]. 光学 精密工程, 2018, 26(8): 2112-2121.

[4] Cui Z J, An J S, Cui T S. Real-time and anti-occlusion visual tracking algorithm based on multi-layer deep convolutional features[J]. Acta Optica Sinica, 2019, 39(7): 0715002.
崔洲涓, 安军社, 崔天舒. 基于多层深度卷积特征的抗遮挡实时跟踪算法[J]. 光学学报, 2019, 39(7): 0715002.

[5] Dong Q J, He X D, Ge H Y, et al. Adaptive merging complementary learners for visual tracking based on probabilistic model[J]. Laser & Optoelectronics Progress, 2019, 56(16): 161505.
董秋杰, 何雪东, 葛海燕, 等. 基于概率模型的自适应融合互补学习跟踪算法[J]. 激光与光电子学进展, 2019, 56(16): 161505.

[6] Wang L J, Ouyang W L, Wang X G, et al. Visual tracking with fully convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3119-3127.

[7] Danelljan M, Hager G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]// 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 621-629.

[8] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4293-4302.

[9] Held D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 749-765.

[10] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1420-1429.

[11] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[M]// Hua G, Jégou H. Computer vision-ECCV 2016 Workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[12] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[13] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 445-461.

[14] Guo Q, Feng W, Zhou C, et al. Learning dynamic Siamese network for visual object tracking[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 1781-1789.

[15] Ma C, Huang J B, Yang X K, et al. Hierarchical convolutional features for visual tracking[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3074-3082.

[16] Liu Q, Yuan D, He Z Y. Thermal infrared object tracking via siamese convolutional neural networks[C]// 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), December 15-17, 2017, Shenzhen, China. New York: IEEE, 2017: 17614033.

[17] Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]// 2012 IEEE Conference on Computer Vision and

Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 1822-1829.

[18] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422.

[19] Zhang J M, Ma S G, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization[M] // Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8694: 188-203.

[20] Hong Z B, Chen Z, Wang C H, et al. MUlti-Store Tracker (MUSTer): a cognitive psychology inspired approach to object tracking[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 749-758.

[21] Ross D A, Lim J, Lin R S, et al. Incremental learning for robust visual tracking[J]. International Journal of Computer Vision, 2008, 77(1/2/3): 125-141.